Technical Report GriPhyN-2001-xx
www.griphyn.org

DRAFT: COMMENTS AND MATERIAL SOLICITED

GriPhyN-LIGO Project Plan for Year 2

December 22, 2001

Developed by members of the GriPhyN-LIGO Project Team

Submit changes and material to: Ewa Deelman, editor
deelman@isi.edu

1 Project Year 2: October 1, 2001 to September 30, 2002

In the first year and the first quarter of Year 2, we focused on basic Virtual Data aspects, such as materializing data based on a user's request.

We have constructed a prototype, which was shown at the SC 2001 conference in Denver. This demo exposed much of the GriPhyN infrastructure. A Virtual Data Request was formed by a broker on behalf of the client; this XML document was sent to the Planner, which planned the necessary computations and data movements based on the resources available at Caltech, UWM and the SC showfloor. The plan was then sent to CondorG for execution, computation was initiated and monitored, and the result delivered to the client.

We also demonstrated the feasibility of the integration of Grid middleware, such as Globus and CondorG with the existing LIGO Data Analysis System (LDAS). As part of this integration, Globus services such as the Globus Resource Allocation Manager (GRAM) were interfaced to the LDAS job submission system. Functionality was also provided to enable the staging of data in and out of LDAS via the GridFTP protocol. As a result, the Globus Security Infrastructure was provided as a secure way of accessing the LDAS compute resources at Caltech and UWM.

In the second year of the project, we will focus our efforts in four directions:

- Increase the complexity of Virtual Data requests and increase the amount of available Virtual Data.
- Further investigation of Virtual Data concepts, including the evaluation and implementation of the Transformation Catalog.
- Request planning, investigate available methodology and examine the fault tolerance requirements.
- Evaluate the use of Globus Grid Security Infrastructure (GSI) and the newly developed Community Authorization Service (CAS) in the LIGO environment.

A unifying framework for this work is the Search for Continuous Wave (or Pulsar) Sources of Gravitational Waves which has been posed in the LIGO collaboration. This challenge is focused on large-scale pulsar search and is described in the next section.

2 The Pulsar Search Mock Data Challenge

The Pulsar search takes 1-D time series strain data (which is termed the gravitational wave channel) and creates 2-D frequency-time (f-t) maps or images. These are then processed using digital signal processing techniques to look for evidence of weak, continuous-wave (CW), monochromatic spectral features. The search is conducted in several stages.

First, the gravitational wave channel is extracted from full frame files. This channel is corrected for dropouts, instrumental response function (the calibration of the channel from ADC counts to physical units), and down-sampled from the original 16.384 kHz to 2.048 kHz. Then 1800 1-second, single channel frames are concatenated to form 30 minute duration frames. In order to build a time-frequency image, Short Fourier Transforms (SFTs) are created on the data using code that uses LDAS. Each resulting frame is now a composite of many SFTs that correspond to the original 30 minutes of time series data. These processed frame files are approximately 10 Mbytes in size. The SFT frames can be further combined to construct time-frequency images that refer to about 6 months of data over a narrow frequency range.

This processing requirement fits well into the GriPhyN Virtual Data Grid model. The basic elements of the GriPhyn VDT can be used to materialize these derived data products, based on user requests. For example, if a scientist wishes to search for a known pulsar in the data, he would request a 2-D image spanning the narrow frequency band (bandwidth of a ~few x 10^{-3} Hz.) for a range of time, and build the corresponding virtual data request. This request would then be serviced by the GriPhyN infrastructure.

Although we addressed issues of channel extraction and transpose in the current prototype this Mock Data Challenge brings with it new challenges. We now have a wider range of derived data products, some of them in the frequency domain.

3 Virtual Data Requests

Our goal in Year 2 is to expand on the base system developed in Year 1 by continuing research on scientific Virtual Data for LIGO, extending the scientific complexity and sophistication of the Virtual Data products that GriPhyN can provide, as well as the quantity of data that is accessible by Grid tools. In order to accomplish that, we need to address the following issues:

Grid-enable the scientific analysis of pulsar application codes. Continue the development of the Globus GRAM interface to the LIGO Data Analysis System (LDAS), with an emphasis on registration of virtual data products with the LIGO event monitor database. LIGO data replication to Tier II sites using the Globus Replica Catalog tools and CondorG, along with registration of replicated data and subsequent virtual data (since Tier II sites may post-process raw data before archiving). In order to give the GriPhyN-LIGO Virtual Data (GLVD) access to all LIGO data, it needs to be able to see the catalog of such data. As the LDAS DB2 database becomes synchronized with the contents of the HPSS and LHO/LLO holdings, so the Replica/Derived-data catalogs should mirror it.

Specific milestones are as follows:

Q1: [UWM] Prototype a Globus/LDAS interface, which uses GSI security. (done)

Q2: [UWM] Broaden the GRAM/LDAS interface to accommodate a greater variability and functionality of LDAS Virtual Data Products, such as SFTs, concatenation, decimation and resampling.

Q3: [CIT, USC, UWM] Design a Data Discovery mechanism for discovery of data replicas on a Grid. One focus of this work is to take into account the ability to interact with the LDAS Diskcache resources in order to enable external visibility of LIGO/LDAS data on the Grid.

Q4: [CIT, USC, UWM] Implementation of the Data Discovery mechanism to support the pulsar search.

4 Virtual Data Concepts

So far, we have explored only a very small subset of LIGO's Virtual Data space. In the second year, we plan to expand this further by providing support for a larger number of transformations used to create Virtual Data products. In the first quarter of year 2, we have proposed a design for the Transformation Catalog (TC), which analogous to the Replica Catalog, performs a mapping from a logical space to the physical space. In the TC, logical transformation names are mapped to physical instances of the transformations, indicating the support OS platform, configuration files needed, etc… Initially, we will populate the catalog with the transformations needed to support the pulsar search. We expect that implementing and experimenting with the TC will allow us to better understand Virtual Data concepts.

A pulsar search is a prime candidate for the application of virtual data concepts and grid based computational tools. This is because there are many possible transformations that one can perform on these data, each specified by a putative source location on the celestial sphere and by parameters that characterize source's intrinsic properties. The entire data set for a one-year search might be of order: (2048 samples/sec) x (2 bytes/sample) x (3x10^7 sec) = 120 Gbytes. This sample set covers a 1kHz frequency bandwidth.

A given pulsar will emit in a much narrower frequency range. Frequency variations arise from modulation due to the earth's motion around the sun, resulting in approximately a frequency variation of df/f ~ 10^{-4}. Consequently, a year of data can be searched for a 1 kHz pulsar by using only (10^{-4}) x 120GBytes = 12 MBytes of data. A reasonable chunk of data to process at once is 50 Mbytes. A priori, though, it is not known *which* 50 Mbyte subset of the data to use.

A simple strategy for grid-enabling a wide-area pulsar search is as follows.

- A server distributes ~50 MBytes of data to a large number of clients

- Each client searches this 50 MBytes of data for pulsars located anywhere on the sky (or in some specific subregion)

- The client can search in principle about 10^{14} sky locations, so the client can compute "forever" based on this small data volume.

- Each time that a candidate source is found at a given sky location and set of intrinsic source parameters, the relevant parameters are returned to the server.

- Statistical studies and more extensive follow-up analysis are done on the candidate sources. One of the LIGO collaboration groups at the Albert Einstein Institute in Potsdam, Germany will be producing a standalone code that implements this strategy, for use on a dedicated computing cluster.

The immediate task will be migrating and adapting this code to one that can use the entire iVDGL and GriPhyN computing grid.

Milestones:

Q1: [USC] Design the Transformation Catalog. (done)

Q2: [USC] Implement the Transformation Catalog.

Q3: [CIT, USC, UWM] Explore the design of the Derived Data Catalog, which specifies how Virtual Data products are materialized.

Q3: [USC,UC, ANL] Unification of the catalog schemes used by CMS and LIGO – basing it on a common VDT 2.0 release.

Q4: [CIT, UWM] Apply replication concepts by developing a real-time international mirror, and a fault-tolerance replica at UW-Mil. This will help us clarify what are the issues in performing bulk operation in the Virtual Data Grid environment Use of the Transformation Catalog to materialize Virtual data required in the pulsar search.

5 Request Planning and Fault Tolerance

So far, planning has been performed on a very small scale. In year 2, we will evaluate the existing planning techniques and determine their applicability to the GriPhyN planning problem. Our Planner will need to take into account a large number of resources and initially focus on processing single requests. In the future, we will also explore bulk request processing.

Milestones:

Q2: [USC] Specify the planning requirements.

Q3: [USC] Evaluate the available planning software.

Q4: [USC] Prototype a reasonably sophisticated planner for GLVDG.

Another area not addressed yet is the problem of fault tolerance. Here, we need to perform the following tasks:

- Specify fault tolerance requirements for LIGO and for GriPhyN in general.
- Assess existing fault and failure issues within the LIGO application.
- Assess the applicability of existing fault tolerance analysis techniques for distributed systems to GriPhyN/LIGO.
- Explore approaches for producing a fault tolerant DAGMan.

As part of our fault tolerance analysis, we will define a framework for fault tolerance analysis within the context of GriPhyN, at several levels: system, task and DAGMan levels.

We will also perform a design and sensitivity analysis by:

- Exploring fault tolerance issues of alternative designs.
- Exploring sensitivity of alternative designs to various failures.

Finally, we will evaluate techniques for performing testing and diagnostics, with emphasis on fault/failure detection and recovery.

Milestones:

Q2: [USC] Specify fault tolerant requirements. Assess existing fault/failure issues within LIGO.

Q3: [USC] Define framework for fault tolerant analysis. Assess the applicability of existing techniques to GriPhyN/LIGO. Explore approaches for producing a fault tolerant DAGMan.

Q4: [USC] Recommend an approach for the incorporation of fault tolerance in LIGO.

6 Security

LIGO has sophisticated requirements where different groups need to access different parts of the data stream. While the Globus model can deal with this, GriPhyN would like to stop short of actually implementing it, but rather act in a consulting role, with LIGO or CACR staff doing the implementation. As part of the collaboration, we will be architecting the security infrastructure, including questions of certificate authority and revocation, interaction with policy committees, delegation of authority. We will evaluate the applicability of the CAS in access control management.

To evaluate the security infrastructure, we will build a parallel, shadow network of gateways into existing LDAS installations. These will be implemented using secure Globus toolkit. Once this network is shown to be working and debugged, we will gradually migrate onto this new fabric and away from our current security solution. (Procedure analogous to how a new freeway is built next to existing highway without disrupting service at any time).

Milestones:

Q1: [UWM] Prototype of basic Grid Security Infrastructure for LDAS: GRAM and GridFTP interface to LDAS. (done)

Q2: [CIT, USC, UWM] Assess LIGO's Virtual Data Grid security needs. Evaluate the applicability of CAS to LIGO's access control needs.

Q3: [CIT, UWM] Implement enhanced security features on existing LDAS installations.

Q4: [CIT, USC, UWM] Demonstration of secure access to multiple LDAS sites.

7 Project Year 2 Milestone Summary By Quarter

Y2 Q1: Oct-Dec 2001:

Demonstration of Virtual Data (GLVD) with a small amount of data in replica catalog, and a small number of channels. (presented at SC'01, November 2001)

Y2 Q2: Jan-Mar 2002:

Building up GLVD prototype to wider quantity of Virtual Data. Implementation of the Transformation Catalog. Assess GLVD security needs, with specific emphasis on CAS.

Y2 Q3:Mar-Jun 2002:

Specification of planning and fault-tolerance requirements. Development of data discovery mechanisms. Implementation of Grid-based security.

Y2 Q4: July-Sept. 2002

Demonstration of Virtual Data technologies developed in Year 2, to be conducted at SC'02. The prototype will be based on the Pulsar search and will include a broad range of Virtual Data products.